Jahangirnagar University
Department of Statistics and Data Science
Part IV B. Sc (Hons.) Final Examination-2022

Course Name	:	Statistical Inference II	**Course Code**	:	STAT-401
Total Marks	:	70	**Time**	:	4 hours

Answer any ***FIVE (05)*** of the following questions. All questions carry equal marks

1. a) What do you mean by equivalence estimator? Define the location and scale invariance of an estimator, minimum risk equivalence estimator, and Bayesian estimator.

b) What do you mean by Pitman Closer and Pitman Closest Estimator? What are the locations and scale parameters?

c) Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the density,

$$f(x;\theta) = \frac{1}{\theta};0 < x < 20$$

Find the Pitman estimator for scale parameter θ.

2. a) What is the ellipsoid of concentration and Wilk's generalized variance?

b) Discuss and derive Bhattacharyya's lower bound. Discuss Chapman Robbins and Kiefer inequality with suitable examples.

c) What is joint completeness? Let $X_1, X_2, \ldots, X_n$ be a random sample from,

$$f(x;\ \theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} I_{(\theta_1,\theta_2)}(x)$$

where, $\theta_1 < \theta_2$. Let $Y_1 = \min[X_1, X_2, \ldots, X_n]$ and $Y_n = \max[X_1, X_2, \ldots, X_n]$. Show that Y_1 and Y_n are completely sufficient.

3. a) In what situations might bootstrapping be preferred over traditional parametric methods for statistical inference? Describe the Bootstrap algorithm for estimating standard errors.

b) Given the data set: 10, 15, 20, 25, 30, use the Jackknife method to estimate the variance. Test whether the Jackknife estimator of variance is unbiased based on this data. Also, construct a 95% confidence interval for the Jackknife estimator of variance.

c) Distinguish between Confidence Interval and Credible interval. Let $x_1, x_2, \cdots, x_n$ be a random sample from $N(0, \sigma^2)$. Find a large sample confidence interval for σ^2 with an approximate confidence coefficient $1-\alpha$.

4. a) Define Bayes factor. When should one use a Bayes factor instead of p-values for hypothesis testing?

b) Show that Bayes factor in favor of alternative hypothesis is nothing but the likelihood ratio of alternative and null hypothesis.

c) Suppose $X \sim N(\theta, 2)$. Assume that H_0, H_1 are equally likely. To test $H_0: \theta = 0$ v.s. $H_1: \theta = 2$,

   i. find Bayes factor in favor of H_1 and interpret the result.

   ii. If $n = 10, \bar{x} = 1$, obtain Posterior odds in favor of H_0.

5. a) What is confidence set, confidence belt and confidence interval.

b) Define uniformly most accurate unbiased (UMAU) family of confidence set with example.

c) Imagine a factory where the time taken by a machine to produce a widget is believed to be exponentially distributed with a parameter δ. After observing the machine's performance, it's deduced that the time t to produce a widget is governed by:

$$f(t \mid \delta) = 3\delta e^{-3\delta t}; \quad t>0$$

From previous data, the factory manager believes that the parameter δ itself has an exponential distribution with a parameter of 4. Given a recorded production time t, what is the posterior distribution of δ?

6. a) What do you mean by Conjugate family of distribution? Show that family of univariate normal distribution is closed under multiplication.

b) Explain the concept of a conjugate prior in Bayesian statistics. Provide an example of a likelihood function and find its conjugate prior, interpret how the conjugate prior simplifies the process of updating beliefs with new data.

c) Define Jaffrey's' Non-Informative prior. Show that Jeffrey's prior of Bernoulli distribution with parameter θ, is nothing but a proper Beta (1/2,1/2) Distribution.

7. a) Define Most Powerful Test (MP), Uniformly Most Powerful Test (UMP), Unbiased Test. If UMP test exists, discuss any technique or theorem to find it.

b) Let $X_1, X_2, ..., X_n$ be a random sample from $N(0,\theta)$, where θ is unknown. Show that, there is no uniformly powerful test for testing the simple hypothesis $H_0 : \theta = \theta'$, where θ' is a fixed number against the alternative composite hypothesis $H_1 : \theta \neq \theta'$.

c) Let X have the PMF $f(x;\theta) = \theta^x (1-\theta)^{1-x}$, $x = 0,1$, zero elsewhere. To test the hypothesis $H_0 : \theta = \frac{1}{4}$ vs. $H_1 : \theta < \frac{1}{4}$ by taking a random sample of size 10, show that, the best critical region is $\sum_{i=1}^{10} x_i \leq 1$ and also a UMP Test. Also, find the power function $\gamma(\theta)$, $0 < \theta \leq \frac{1}{4}$ and draw the power curve.

8. a) What do you mean by Wald's Sequential Probability Ratio test? When would you choose a sequential sampling approach over classical fixed-sample approach for hypothesis testing?

b) Let α_α and β_α be preassigned proper fractions in SPRT defined by $k_0 = \frac{\alpha_\alpha}{1-\beta_\alpha}$, $k_1 = \frac{1-\alpha_\alpha}{\beta_\alpha}$,

show that $\alpha \leq \frac{\alpha_\alpha}{1-\beta_\alpha}$, $\beta \leq \frac{\beta_\alpha}{1-\alpha_\alpha}$.

c) Suppose, $X \sim Poisson(\theta)$. Find the sequential probability ratio test for testing $H_0 : \theta = 0.02$ against $H_1 : \theta = 0.07$. Show that this test can be based upon the statistic $\sum_{i=1}^{n} X_i$. If $\alpha_\alpha = 0.20$ and $\beta_\alpha = 0.10$, find $c_0(n)$ and $c_1(n)$.

২৯২২০৭

Jahangirnagar University
Department of Statistics and Data Science
Part IV B. Sc. (Hons.) Final Examination 2022

Course Title:	Multivariate Analysis	**Course No.:**	STAT-402
Time:	4 hours	**Full Marks:**	70

[Answer any five (05) questions. All questions carry equal marks]

1. a) What do you mean by multivariate analysis? How can researchers use multivariate analysis to make a productive analysis plan?

b) Let $X_1, X_2, \dots, X_n$ are independent and identically distributed (IID) with $\mathbf{N_p(0, \Sigma)}$. Find the maximum likelihood estimate (MLE) of $\mathbf{\Sigma}$. Show that it is an unbiased estimator of $\mathbf{\Sigma}$.

c) Let $\mathbf{X}$ be $\mathbf{N_3}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = [5 \quad 3 \quad 2]$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 9 \end{bmatrix}$.

(i) Are $\frac{X_1}{2}$ and $X_1 + X_3$ independently distributed? Explain.

(ii) Find the conditional distribution of X_1 given $X_1 - 2X_2 + \frac{X_3}{3}$.

2. a) How can you access the assumption of multivariate normality? Describe. Discuss the steps of detecting outliers in a higher dimension of the multivariate data set.

b) If $X_1, X_2, \dots, X_n$ be independent observations from a population with mean μ and finite covariance Σ, then show that $n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu)$ is approximately distributed as χ^2_p for large $n - p$.

c) What are the most common multivariate quality control charts? Define them with their uses. *Construct a* T^2 *chart for the data in Table* 1. *Use* $\alpha = 0.01$. *Hence , comment.*

Table 1: Two measurements of stiffness with bending strength.

x_1	1232	1115	2205	1897
x_2	4175	6652	7612	10914

Where, $x_1 =$ stiffness and $x_2 =$bending strength are two measurements in pounds/$(inches)^2$ for a sample of 4 pieces of a particular grade of lumbers.

3. a) Explain the use of distance in multivariate analysis. Which distance is more preferable in statistics and why?

b) Define Hoteling $\mathbf{T^2}$ statistics with its applications. Let $X_1, X_2, \dots, X_n$ be a random sample from $\mathbf{N_p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then show that $\mathbf{T^2}$ can be expressed as a function of Wilk's Lambda.

c) What do you mean by confidence region of μ, where $X \sim N_p(\mu, \Sigma)$. Construct the 95% confidence region of μ for n=40 pairs of observations having,

$$\bar{x} = \begin{bmatrix} 0.567 \\ 0.603 \end{bmatrix} \; and \; S = \begin{bmatrix} 0.014 & 0.012 \\ 0.012 & 0.015 \end{bmatrix}$$

Hence, check whether, $\boldsymbol{\mu}' = [0.560 \quad 0.580]$ is in that confidence region.

4. a) Define the multivariate multiple regression model. What are the advantages of the multivariate multiple regression model than its counterparts?

b) Suppose the maximum likelihood estimator of $\boldsymbol{\beta} = [\boldsymbol{\beta}_{(1)} \vdots \boldsymbol{\beta}_{(2)} \vdots \cdots \vdots \boldsymbol{\beta}_{(m)}]$ is $\hat{\boldsymbol{\beta}} = \mathbf{(Z'Z)^{-1}Z'Y}$ determined under the multivariate multiple regression model with the errors ε have a normal distribution and the design matrix $\mathbf{Z}$ having full rank $\mathbf{(Z) = r + 1, n \geq (r + 1) + m}$, then show that $\hat{\boldsymbol{\beta}}$ has a normal distribution with $\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathbf{Cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) = \boldsymbol{\sigma}_{\mathbf{ik}}\mathbf{(Z'Z)^{-1}}$; $\mathbf{i, k = 1, \dots, m}$.

c) Suppose 10 American companies' sales and profit are two responses regressed on the covariate asset, specify the mathematical form of the multivariate multiple regression model. How could you test $\mathbf{H_0}: \boldsymbol{\beta}_2 = 0$ for the model.

5. a) What is principal component analysis (PCA)? Describe the advantages and disadvantages of PCA. Also explain the role of PCA in multivariate analysis with suitable example.

b) Let $\mathbf{X}' = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p]$ have covariance matrix Σ, with eigen value-eigen vector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \cdots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$. Let $\mathbf{Y}_1 = \mathbf{e}_1'\mathbf{X}, \mathbf{Y}_2 = \mathbf{e}_2'\mathbf{X}, \cdots, \mathbf{Y} = \mathbf{e}_p'\mathbf{X}$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \text{var}(\mathbf{X}_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{var}\cancel{(X_i)}(Y_i)$$

c) Determine the population principal component $\mathbf{Y}_1$and $\mathbf{Y}_2$ for the covariance matrix

$$\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix} \qquad \lambda_1 -$$

Also calculate the proportion of total population covariance explained by the first principal component.

6. a) Define canonical correlation and canonical variables. What is the purpose of performing the canonical correlation analysis? State and prove all the properties of canonical correlation analysis.

b) Suppose $p \leq q$ and $\rho_1^{*2} \geq \rho_2^{*2} \geq \ldots \geq \rho_p^{*2}$ be the p ordered eigenvalues of $\Sigma_{11}^{-1/2}\,\Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21}\,\Sigma_{11}^{-1/2}$ with corresponding eigenvectors $e_1, e_2, \ldots, e_p$. Also let $f_1, f_2, \ldots, f_p$ be the eigen vectors of $\Sigma_{22}^{-1/2}\,\Sigma_{21}\,\Sigma_{11}^{-1}\,\Sigma_{12}\,\Sigma_{22}^{-1/2}$, where each f_i is proportional to $\Sigma_{22}^{-1/2}\,\Sigma_{21}\,\Sigma_{11}^{-1}e_i$. The random vector $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ have covariates Σ_{11} and Σ_{22}, respectively with $\text{Cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \Sigma_{12}$. Then show that the k^{th} pair of canonical variates, $k = 1,2, \ldots, p$, $\mathbf{U}_k = \mathbf{e}_k'\Sigma_{11}^{-1/2}\mathbf{X}^{(1)}$ and $\mathbf{V}_k = \mathbf{f}_k'\Sigma_{22}^{-1/2}\mathbf{X}^{(2)}$ maximizes $\text{Corr}(\mathbf{U}_k, \mathbf{V}_k) = \rho_k^*$

c) Use Bartlett test to find the number of significant canonical correlations for the following studies,

(i) With $n = 456, p = 4, q = 5, \widehat{\rho}_1^* = 0.67, \widehat{\rho}_2^* = 0.23, \widehat{\rho}_3^* = 0.07$ and $\widehat{\rho}_4^* = 0.03$.

(ii) With $n = 744, p = 6, q = 4, \widehat{\rho}_1^* = 0.46, \widehat{\rho}_2^* = 0.38, \widehat{\rho}_3^* = 0.12$ and $\widehat{\rho}_4^* = 0.04$.

7. a) What do you mean by discrimination and classification in multivariate analysis? What are the goals of discrimination and classification?

b) Discuss Fisher's linear discrimination function for two multivariate normal populations. Also, describe the allocation rule based on this function.

c) Consider the two data sets, $x_1 = \begin{bmatrix} 37 \\ 24 \\ 47 \end{bmatrix}_{3\times2}$ and $x_2 = \begin{bmatrix} 69 \\ 57 \\ 48 \end{bmatrix}_{3\times2}$ $\quad x_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix} \quad x_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$

For which, $\bar{x}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \bar{x}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$ and $S_{pooled} = \begin{bmatrix} 11 \\ 12 \end{bmatrix}$ $\quad S_p = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}_{2\times2}$

(i) Calculate linear discrimination function

(ii) Classify the observation $x_0' = [27]$ as population Π_1 or population Π_2 using any suitable rule.

8. a) What do you mean by cluster analysis? When should it be used instead of factor analysis? What are the different distance and similarity measures?

b) What are the different types of clustering algorithm? Describe them in brief.

c) Consider the matrix of distances

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & & & \\ 2 & 1 & 0 & & \\ 3 & 11 & 2 & 0 & \\ 4 & 5 & 3 & 4 & 0 \end{array}$$

Cluster the four items using each of the following procedures.

(i) Single linkage hierarchical procedure.

(ii) Complete linkage hierarchical procedure.

(iii) Average linkage hierarchical procedure.

(iv) Draw the dendrograms and compare the results in (i), (ii), and (iii).

১২২

Jahangirnagar University
Department of Statistics and Data Science
Part IV B. Sc. (Hons.) Final Examination 2022

Course Title: Design of Experiment **Course No.:** STAT-403
Time: 4 hours **Full Marks:** 70

[Answer any five (05) questions. All questions carry equal marks]

1. a. Explain the term (i) Treatment (ii) Experimental unit (iii) Yield (iv) Block (v) Experimental error.

   b. Describe the concept of all basic principles of the Design of experiment. Write down how these principles are applied to provide valid interpretations of data.

   c. To compare the four mixtures, three different samples of propellant are prepared from each mixture and readied for testing. Each of the three investigators is randomly assigned one sample of each of the four mixtures and asked to measure the propellant thrust. These data are summarized next.

Mixtures	Investigator		
	1	2	3
1	2,340	2,355	2,362
2	2,658	2,650	2,665
3	2,449	2,458	2560
4	2,403	2,410	2,418

   i. Identify the blocks and treatments for this experimental design.
   ii. Indicate the method of randomization.
   iii. Write down the name of the appropriate design for this data. Explain the reason for your choice.
   iv. Conduct the ANOVA table for the data set when the observation for 3rd mixtures and 2nd investigator is missing.

2. a. Define variance component analysis. Write a random effect model and its assumption for two-way Classification with a single observation per cell. Conduct an ANOVA table and estimate the variance components. Find the variance of the estimates.

   b. The management of a poultry farm collected 10 varieties of dry concentrate for chicks of 7 different ages. Each concentrate was continued up to the age 45 days of chicks concentrate to the chicks and 5 varieties of dry concentrate and 4 different ages were randomly selected. The concentrates and different ages were given and then the weights of the chicks were recorded. Based on this information complete the following table and find the variance of the estimates

SV	df	SS	E(MS)	F
Age	-	5.41	-	-
Concentrate	-	2.17	-	-
Error	-	0.55	-	
Total	-	-		

3. a. Define Latin square design (LSD). Write down the reason why Latin square design is called incomplete block design. Make a comparative study between LSD and Graco LSD.

   b. Conduct a layout plan for the Latin square design (i.e.,) LSD with *k treatment*, state, and explain the model of this design with necessary assumptions. Set up the ANOVA table and write down the decision rules for testing different hypotheses. Find the efficiency of this design compared to CRD.

4. a. What do you mean by analysis of covariance (ANCOVA)? Give an example where it is used.

b. Set up a mathematical model for ANCOVA in RBD with two concomitant variables and discuss the analysis procedure of such data.

5. a. What is split plot design? Discuss how it differs from randomized block design and confounded design. Mention the situation where a split-plot design is used.

b. Discuss the procedure of analyzing data obtained from a split-plot design with two factors. Prepare the ANOVA table.

6. a. What is confounding and what ac its necessity? Discuss different types of confounding with examples.

b. Discuss the block consist of 2^4-factorial experiment if ABCD and AC interactions are simultaneously confounded in the same replication. Discuss the procedure of analysis of data to test the hypothesis.

7. a. Define an incomplete block design. When incomplete block design becomes balanced. For a BIBD with the usual parameters, show that

i. $\lambda(v-1) = r(k-1)$

ii. $b \geq v$

iii. $r \geq k$

b. Describe the procedure of analysis of data obtained from a BIB design. Construct a layout plan for a BIB design having parameters b=v =11, r = k = 5, $\lambda = 2$.

8. a. Define Concomitant variables as well as analysis of covariance with examples. Give an example where it is used. Write down the difference between the analysis of variance and the *analysis of covariance.*

b. A feeding trial experiment was conducted to compare the effects of two different feeds on the weight gain of goats in a firm. Feeds were given to selected goats for 3 months and gain weight was recorded. Tabulated results of the experiment showing initial weight *X* (in kg) and gain in weight *Y* (in kg)

	F_1		F_2	
	x	y	x	y
	5.5	1	7.5	2.5
	6.5	2	5	1.5
	4.5	1	6.5	1.8
	7.5	1.6	6	2.0
Total	24	5.6	25	7.8

i. Write down the appropriate ANCOVA model for this data. Justify your choice.

ii. Complete the ANCOVA table.

iii. Conduct a test procedure to test whether two foods are the same.

102207

Jahangirnagar University
Department of Statistics and Data Science
Part IV B. Sc. (Hons.) Final Examination 2022

Course Title: Sampling Technique II Course No.: STAT-404
Time: 2.5 hours Full Marks: 35

[Answer any three (03) questions. All questions carry equal marks]

1. (a) What is the major objective of using probability proportion to size (PPS) sampling? Explain in brief. List a few situations where probability to size with replacement (PPSWR) sampling can be used in actual practice. Also, discuss at least one method of selecting a sample by PPSWR sampling.

(b) Let there be a population of N units and we want to select a sample of n units. For this selection, the population is randomly divided into n groups of sizes $N_1, N_2, \dots, N_n$ such that $\sum_{i=1}^{n} N_i = N$. For the first group, we select N_1 units out of N units with SRSWOR sampling. Then the second group, select N_2 units out of $(N - N_1)$ units with the same sampling scheme and so on. Then one unit is drawn from each group with probability proportion to size of the units in that group. For this schema, show that (under usual notation) an unbiased estimator of population total is given by

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\left(\frac{P_i}{\tau_i}\right)}, \qquad \tau_i = \sum_{j\epsilon N_i} P_{ij}$$

Also, give an expression for the unbiased estimator of $V(\hat{Y})$.

(c) Find the gain in efficiency due to sampling with varying probability over simple random sampling with replacement (SRSWR).

2. (a) Describe a situation where two-stage sampling is appropriate. Explain the difference between stratified sampling and two-stage sampling.

(b) Prove that the mean per second-stage unit in the sample is an unbiased estimate of the population mean for two-stage with equal first-stage units. Also, obtain the variance of the estimate of the mean.

(c) Find the best possible first-stage units in two-stage sampling.

3. (a) Under what circumstances successive sampling is used? Give example.

(b) What are the reasons for using sampling over successive occasions?

(c) Suggest a best linear estimator of a population characteristic (e.g. mean) on current occasion. Verify your proposal. Find the optimum size of unmatched sample on the second occasion.

4. (a) What do you mean by double sampling? Explain the situation where double sampling is necessary compared to other sampling designs. What are the negative features of double sampling? Explain them.

(b) If the sample is random and size n', the second sample is a random subsample of the first size $n_h = \nu_h n'_h$, where $0 < \nu_h < 1$ and ν_h are fixed, then show that $\bar{y}_{st}$ population mean in case of double sampling for regression method of estimation. Show that,
$v(\bar{y}_{st}) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \sum_h^L \frac{W_h S'_h}{n'}\left(\frac{1}{\nu_h} - 1\right)$, where S^2 is the population variance.

5. (a) What do you mean by single-stage cluster sampling with unequal cluster size? Let M_i be the number of elements of the ith cluster. Draw a sample of n clusters with probabilities proportion to their sizes M_i and with replacement.

(b) Suppose that a sample of n cluster is drawn with probabilities $z_i = \frac{M_i}{M_o}$, where $M_o = \sum_{i=1}^{n} M_i$ and with replacement. Show that, an unbiased estimator of population total Y is $\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{z_i}$. Also, find the sampling variance of $\hat{Y}$.

Jahangirnagar University
Department of Statistics and Data Science
Part IV B.Sc. (Hons.) Final Examination 2022

Course Title: Data Mining **Course Code:** STAT-405
Time: 2.5 Hour **Full Marks:** 35

[Answer any **three (03)** of the following questions and all question carries equal marks.]

Q1. (a) What is meant by Data Mining? Briefly mention the Data Mining development and explain the different data mining tasks with examples.

(b) Explain the different types of Knowledge Discovery in Database (KDD). What are the different types of data in Data Mining? Briefly explain with an example.

(c) Briefly discuss the different types of Similarity Measures used in data mining. Is Correlation a similarity measure? Why or why not? Explain.

Q2. (a) Define naïve Bayes classification, and what makes it "naïve" in its assumption?

(b) The following data describes four characteristics from the all-electronics customer database

ID	Age	Income	Student	Credit_rating
1	Youth	High	No	Fair ✓
2	Youth	High	No	Excellent
3	Middle-aged	High	No	Fair
4	Senior	Low	No	Fair ✓
5	Senior	Low	Yes	Fair ✓
6	Middle-aged	Low	Yes	Excellent
7	Senior	Medium	Yes	Excellent
8	Youth	Medium	No	Fair ✓
9	Senior	High	No	Fair
10	Youth	Medium	Yes	Excellent
11	Middle-aged	High	Yes	Excellent
12	Senior	Low	No	Fair
13	Youth	High	No	Fair ✓

Calculate the following terms from the above dataset using income and student: Information gain, gain ratio, and Gini index. consider creadit rating as the class label attribute

(c) Briefly explain the concepts of ID3, C4.5, and CART.

Q3. (a) What is meant by Artificial Neural Network (ANN)? Briefly explain the different steps of ANN.

(b) The following result (Table 1) was found for test data after applying the *k*-Nearest Neighbor (KNN) algorithm to the atmospheric data from a region of Bangladesh to classify the rainfall (RAN) [No Rain and Trace (NRT), Light Rain (LTR), Moderate and High Rain (MHR)] based on Temperature (TEM), Dew Point Temperature (DPT), Wind Speed (WIS), Humidity (HUM), and Sea Level Pressure (SLP) for the optimal value of k=9 and Seventy percent observations were used as training data and the rest of data as test data.

Table 1: Confusion matrix for the test data.

		Predicted		
	Category	**LTR**	**MHR**	**NRT**
Actual	**LTR**	65	6	11
	MHR	12	53	5
	NRT	8	0	73

i) What is the actual number of observations?

ii) Find the prediction accuracy rate and error rate for test data.

iii) Obtain the value of Sensitivity, Specificity, and F_1-score for each category LTR, MHR, and NRT.

(c) Describe the backpropagation method of training an artificial neural network.

Q4. (a) What is meant by Association rule? What are the different types of association rule algorithm? Discuss any one algorithm form the association rule algorithms. Also, write down the advantage and disadvantage of this algorithm.

(b) Suppose there are two items, {A, B} where $A \Rightarrow B$ has support of 15% and a confidence 60%. Because these values are high, a typical association rule algorithm probably would deduce this to be a valuable rule. However, if the probability of purchase item B is 70%, then we see that the probability of purchasing B has gone down, presumably because A was purchased.

Find lift, chi-square, all confidence, max confidence, Kulczynski, cosine, correlation and Imbalance ratio. Comment on your result.

(c) A database has five transactions. Let minimum support, $s = 60\%$, and minimum confidence, $\alpha = 80\%$.

TID	*items bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, K, I, E}

Find all frequent itemsets using Apriori.

Q5. (a) What is web mining? What are the three main categories of web mining? Provide a brief explanation of each category.

(b) Apply k-modes clustering algorithm to divide the items into k=3 clusters. You should use the items (ID-2), (ID-4), and (ID-6) for the initial clusters.

ID	1	2	3	4	5	6	7
Gender	M	M	F	F	M	F	M
Product category	Electronic	Clothing	Clothing	Electronic	Clothing	Clothing	Electronic

(c) Briefly explain the text Mining, Spatial Data Mining, and Temporal Data Mining.

Department of Statistics

Part IV B. Sc. (Hon's) Final Examination-2022

Course Code : STAT-406 **Course Title** : Biostatistics and Epidemiology

Full Marks : 70 **Time** : 4 hours

Answer any **Five (05)** questions from the following.

1. Suppose that T is a random variable representing survival time and that T has p.d.f $f_T(t)$ and survival function $S_T(t)$.

   a) Write down the definition of the hazard function $h_T(t)$, and hence show that $h_T(t) = \frac{f_T(t)}{S_T(t)}$ 4

   b) Show that, 4

   $$h_T(t) = -\frac{d}{dt}\log S_T(t)$$

   and hence describe how you would use an estimate of $S_T(t)$ for a particular dataset, to assess whether an exponential (constant hazard) model is appropriate.

   c) The operating time until failure (in hours) of particular LEDs are modelled as observations of a random variable T with hazard function 6

   $$h_T(t) = \beta t^{-\frac{1}{2}}$$

   where $\beta > 0$ is a parameter of the model.

   If $\beta =$ 0.02, calculate:

   i. The probability that an LED operates longer than 400 hours before failure,

   ii. The time by which 90% of LEDs have failed.

2. a) An engineer is investigating the failure times of a particular type of electronic component and proposes to model failure times as observations of a random variable T. The engineer proposes the following 5

   $$h_T(t) = \frac{\beta}{1+t}$$

   for some parameter $\beta > 0$, as the survival function for T

   Give a reason why this function is a valid hazard function. Show that the survival function for T is given by

   $$S_T(t) = \frac{1}{(1+t)^{\beta}}$$

   and find the corresponding density function.

   b) A sample of n such electronic components is tested over a period of P hours, during which time m of the components were observed to fail. The remaining $n - m$ components had not failed at P hours when observation ended. For each failed component, $i = 1,\dots,m$, the observed failure time is denoted by t_i . 6

   Show that the log-likelihood $l(\beta)$ is given by

   $$l(\beta) = m\log\beta - (n-m)\beta\log(P+1) - (\beta+1)\sum_{i=1}^{m}\log(1+t_i)$$

   And find expressions for the maximum likelihood estimate $\hat{\beta}$ of β and its standard error.

   c) Suppose $n = 12, m = 3, P = 1000, t_1 = 200, t_2 = 800$ and $t_3 = 800$. Find $\hat{\beta}$ and an approximate 95% confidence interval for β. 3

3. a) Define the proportional hazard regression model. Derive the likelihood function for this model. Also, show that the Weibull distribution belongs to the proportional hazard family. 6

   b) Explain the estimation method and inference procedure for proportional hazard regression model. 4

   c) What are the advantages of the Cox proportional hazard model over logistic regression? 4

4. In a study investigating a new treatment for chronic reflux disease, times (in months) were recorded between a patient finishing a successful course of treatment and having a relapse (subsequent onset of symptoms). The following data were the recorded times for 15 patients, a + indicating a right-censored observation.

   6+ 12 26+ 37+ 20+ 19+ 6+ 7 22 12 22 19 13+ 21+ 2+

   a) Calculate the Kaplan-Meier estimate for the survival function of the random variable representing time to relapse. Sketch the estimate on a suitable set of axes (a very accurate sketch is not required, but you should label axes and show any points of discontinuity clearly). 6

b) Write down the estimate of the probability of remaining symptom-free for at least 18 months, and use Greenwood's formula to calculate its standard error. Hence calculate a 95% confidence interval for this probability. 4

c) Calculate the Nelson-Aalen estimate for the cumulative hazard function at 18 months. Transform this estimate to a survival-function estimate and compare it to your answer in (b). 4

5. a) What is epidemiology? Discuss the disease etiology with the epidemic triangle model. 3

b) What are the different levels of disease prevention? Describe different types of Epidemiological Surveillance with examples. 6

c) How do you define disease as infectious? Explain the components of the disease infectious process. Describe the disease transmission mode elaborately. 5

6. a) Define cohort in epidemiology. Distinguish between closed cohort study and open cohort study. What do you mean by bias? What are the different types of biases? Explain different types of systematic errors observed in epidemiological studies. 6

b) What is reliability, precision and accuracy? What do you mean by prevalence? What is the relationship between prevalence and the predictive value of a test? Describe with an example. 4

c) An investigator evaluated 100 patients suffering from major depression as confirmed by the attending psychiatrist. The results are as follows: 4

7. a) Write down the differences between incidence and prevalence. Establish the relationship between prevalence and incidence. 5

b) Estimate the Variance and Confidence Interval for the Prevalence with the usual notation. 4

c) Compare the prevalence and incidence rates of the given condition in two different populations: 5

Population A: Initial prevalence of 120, incidence of 15

Population B: Initial prevalence of 80, incidence of 25.

Discuss the implications of these differences.

8. a) What do you understand by incidence density? How does it differ from incidence, explain with an example. 4

b) What do you understand by standardization of disease occurrence? Why it is important in Epidemiology? 4

c) In the study, 100 men with high-fat diets are compared with 100 men who are on a low-fat diet. Both groups start at age 65 and are followed for 10 years. During the follow-up period, 10 men in the high-fat intake group are diagnosed with prostate cancer and 5 men in the low-fat intake group develop prostate cancer. Compute 95% CI for incidence densities 6

2

S(n) = W(n)

Department of Statistics and Data Science
Part IV B. Sc. (Hon's) Final Examination-2022

Course Code : STAT-407 **Course Title** : Advanced Demography
Full Marks : 35 **Time** : 2.5 hours

Answer any Three (03) questions from the following.

Q1. (a) What is demographic transition theory? How many ways it can be classified?
(b) Explain different stages of demographic transition theory.
(c) Can you fit Bangladesh in any one of these stages of transition theory? Explain your answer.

Q2. (a) What is urbanization? How tempo of urbanization differ from the degree of urbanization?
(b) Measure the tempo of urbanization if the annual average rate of change in the population living in urban areas is assumed arithmetic.
(c) Define city size distribution. What can we know from this distribution? If z = 2.25 (the largest city is Dhaka and the smallest city is Mymensingh), what does it mean?

Q3. (a) Is there any difference between parity data and age-specific data? Based on your answers, do explain. Why is parity data used concerning the Gompertz model?
(b) Please use the given data and direct formula (let, $\hat{\beta}$=1.1845) and fill up the empty column from parity data using the Gompertz model. The information is given:

Age group	Index (i)	Average Parity	$Y_s(i)$	Y(i)=?	F(i)=?	f(i)=?
15-19	1	0.28	-0.6913			
20-24	2	1.23	0.0256			
25-29	3	2.18	0.7000			
30-34	4	2.89	1.4787			
35-39	5	3.43	2.6260			
40-44	6	3.93	4.8097			
45-49	7	4.46	-			
Total						

(c) From the cumulative fertility rate and age-specific fertility rate column, explain and comment on the results of both columns.

Q4. (a) What is adult survivorship probability? Also, estimate it for females based on proportions not orphaned.
(b) Based on the given table, calculate the mean age at maternity and weighing factors, including comments.

Age group of respondents	Mother alive	Mother dead	Unknown maternal orphan-hood status	No. of children born in 7 age group of mothers
15-19	5540	448	6	136
20-24	5995	541	10	409
25-29	2886	769	8	485
30-34	1910	852	9	320
35-39	1661	1234	11	259
40-44	1027	1273	7	94
45-49	855	1556	6	50
50-54	369	1243	6	

(c) Also, calculate female survivorship probability by using (b) when age n = 40 with comments.

Q5. (a) Suppose a couple has 3 children (2 sons and 1 daughter) with a wife aged 48. What is the value of TFR, GRR, and NRR with explanations?

(b) You know, there are several proximate determinants of variables of estimating fertility, why does Bongaart use only four variables? Estimate the four indices with comments based on the given values, such as TFR=2.7, TMFR=4.2, u=0.581, e=0.88, i=5.8, and TA=0.18.

(c) Show the impact on fertility reduction and identify the variable that is responsible more for reducing fertility.

Department of Statistics and Data Science
Part IV B. Sc. (Hon's) Final Examination-2022

Course Code : STAT-408 **Course Title** : Stochastic Process
Full Marks : 35 **Time** : 2.5 hours

Answer any **Three (03)** questions from the following.

✓

1. (a) Define the Stochastic Process with a suitable example. What are the different types of Stochastic Process?

   (b) Why Stationarity is important in Stochastic Process? Let $X(t) = A_0 + A_1 t + A_2 t^2$, where $A_i, i = 0,1,2$ are uncorrelated random variables with mean 0 and variance 1. Find the mean value function and the covariance function of $\{X(t), t \in T\}$.

   (c) Define Martingales Process. Let $Z_i; i = 1,2,\ldots$ be a sequence of i.i.d. random variables with $E\{Z_i\} = 1$ and let $X_n = \prod_{i=1}^{n} Z_i$, then show that $\left\{\{X_n\}_{n=1}^{\infty}\right.$ is a martingale.

2. (a) Define the birth and death process with an example. Discuss the relationship between state transition probabilities and birth and death rates.

   (b) State and derive the Kolmogorov's backward differential equation. Also, derive the backward differential equation for the birth and death process.

   (c) Define continuous time Markov chain. How will you find the limiting probabilities of the continuous-time Markov chain? Show that, for a continuous time Markov chain, the rate at which the process leaves any state ***j*** equals the rate at which it enters the state.

3. (a) When is a Counting process said to be a Poisson Process? Find the distribution of the number of arrivals in a Poisson Process.

   (b) If $\{N(t), t \geq 0\}$ is a Poisson Process with rate $\lambda > 0$, then prove that for all $s > 0, t > 0$, $N(s+t) - N(s)$ is a Poisson random variable with mean λt.

   (c) Two types of claims are made to an insurance company. Let $N_i(t)$ denote the number of types i claims made by time t, and suppose that $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent Poisson Processes with rates $\lambda_1 = 10$ and $\lambda_2 = 1$. The amount of successive type 1 claims are independent exponential random variables with a mean of \$1000 whereas the amount from type 2 claims are independent exponential random variables with a mean of \$5000. A claim for \$4000 has just been received; What is the probability it is a type 1 claim?

4. (a) Define the M/M/1 queuing model. State the assumptions of the queuing process. What do we measure in the queuing process?

   (b) For the M/M/1 queuing model, find the distribution of the time an arriving customer spends in the system, X. Hence, find the average amount of time a customer spends in the system.

   (c) Suppose Agrani Bank, JU offers one service counter for the students at the university. Let students arrive at the bank at random at an average of 15 per hour. If the service time has a negative exponential distribution with a mean of 3 minutes; find,

      (i) The average time a student in the queue

      (ii) Average time a student waits before getting service

      (iii) Service utilization rate.

      (iv) Find the probability that an arriving student has to wait more than 3 minutes in the bank.

5. **(a)** Define the Renewal Process. State and derive the distribution of the Renewal Process.

(b) What is Stopping Time? Show that for large t, the number of renewals per unit of time converges to $\frac{1}{\mu}$.

(c) Potential customers arrive at a full-service, one-pump gas station at a Poisson rate of 20 cars per hour. However, customers will only enter the station for gas if there are no more than two cars (including the one currently being attended to) at the pump. Suppose that the amount of time required to service a car is exponentially distributed with a mean of five minutes.

(i) What fraction of the attainment's time will be spent servicing cars?

(ii) What fraction of potential customers are lost?

Department of Statistics and Data Science

Part IV B. Sc. (Hon's) Final Examination-2022

Course Code : STAT-409 **Course Title** : Bioinformatics

Full Marks : 35 **Time** : 2.5 hours

Answer any **Three (03)** questions from the following.

1. (a) What are the important roles of Statistics in the field of Bioinformatics? Distinguish among deoxyribonucleic acid (DNA), Gene and Chromosome. Why DNA is called the carrier of the genetic information? Explain.

(b) Define the following terms:
i) Inheritance pattern, ii) Trait, iii) Mutation, iv) Crossing over, v) Haploid and vi) Diploid

(c) How do two DNA segments differ? Explain the central dogma of a life. With a suitable example, explain all the possible types of genotypes for a diallelic locus.

2. (a) What do you mean by locus, allele, and genotype? Distinguish between genotype and phenotype.

(b) What is pedigree? Write down the characteristics of disease transmission pattern with pedigree:

(i) X-Chromosomal dominant inheritance.

(ii) Y-Chromosomal dominant inheritance.

(c) Describe penetrance function. Acute Lymphocytic Leukemia (ALL) is a rare form of Leukemia. Let the prevalence of roughly 1 in 30,000 persons. Also, assume that the disease transmission follows an autosomal dominant inheritance pattern in a familial study. If the penetrance is reduced with about 85% of individuals carrying the disease-causing allele, calculate the genotypic risk ratio (GRR) and hence interpret.

3. (a) What is a recombination function? Mention the Mendel's law of independent assortment for two genetic loci. How would you apply the penetrance function to model the dominant inheritance pattern for a family with four generations? Explain a diallelic locus.

(b) What is reduced penetrance for an ancestry pattern? Define GRR, and hence introduce its function to model a particular ancestry pattern.

(c) Compare the following two ancestry charts with respect to the characteristics of disease transmission patterns over the generations.

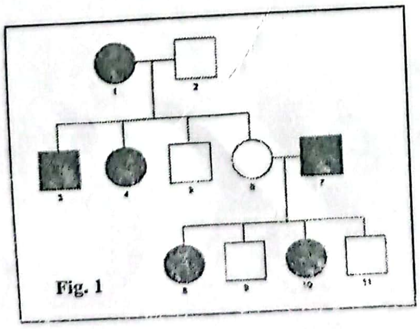


Fig. 1

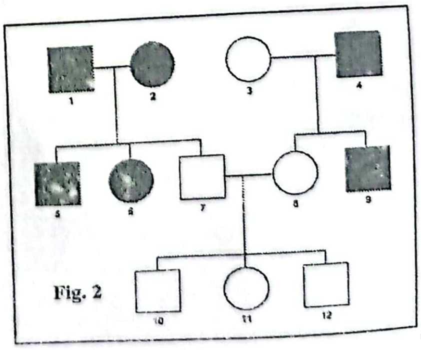


Fig. 2

4. (a) Distinguish between: i) the Similarity and homology of DNA sequences, ii) global and local alignment. What is the multiple sequence alignment (MASA)? Mention some available approaches for applying such alignment.

(b) What is genotyping? Mention some genotyping methods known to you. Classify the mentioned methods according to the feature of "Throughput". How would you select the best one for your analysis from all the genotyping methods available?

(c) Define the Next-generation sequencing (NGS). Write down the important characteristics of NGS.

5. (a) What is a genetic marker? Write down the properties of a genetic marker. How can you check the quality of the genetic markers for your analysis? Explain.

(b) Calculate the heterogeneity of a marker for the following cases,

(i) The major allele frequency is 0.7

(ii) Both the major and minor allele frequencies are the same.

(c) Let, the genotyping of n individuals yield the following genotype frequencies, $P(A_1A_1) = 0.49$, $P(A_1A_2) = 0.42$, $P(A_2A_2) = 0.09$, respectively with alleles A_1 and A_2. What are the allele frequencies if the population is in HWE?

Department of Statistics and Data Science

Part IV B. Sc. (Hon's) Final Examination-2022

Course Code :	STAT-410	**Course Title** :	Categorial Data Analysis
Full Marks :	70	**Time** :	4 hours

Answer any **Five (05)** questions from the following.

1. (a) Mention different types of response variables used in modeling regression type model. Consider a multiple regression model with three parameters The null hypothesis is, $H_0: \beta_1 = \beta_2$. Use the Wald test and the Lagrange Multiplier test to assess this hypothesis.

(b) Suppose you have a dataset following a Poisson distribution. Two models are considered: Model A with one parameter Λ_1 and Model B with two parameters Λ_1 *and* Λ_2. Perform a Likelihood Ratio Test to determine if adding the second parameter in Model B significantly improves the fit compared to Model A.

2. (a) List out the assumptions needed.for measuring Spearman's rank correlation and Kendall's tau-b in the context of ordinal data.

(b) Consider a study that examines the relationship between the frequency of exercise and the perceived level of fitness among a group of individuals. The data consists of ordinal variables. Compute Kendall's tau-b for the association between exercise frequency and fitness perception.

Participants	Exercise (X)	Fitness (Y)
1	2	3
2	1	2
3	3	4
4	3	3
5	2	3
6	3	4

(c) Consider a study that investigates the effectiveness of two different teaching methods, A and B, in improving the pass rates of students in a particular course. Construct a 95% confidence interval for the difference in proportions $(p_A - p_B)$. Interpret the interval in the context of the study. The data is as follows:

Methods A	Students=150	Passed=85
Methods B	Students=120	Passed=80

3. (a) A sample of psychiatric patients by their diagnosis and by whether their treatment their treatment prescribed drugs. Construct and interpret a 95% confidence interval for the population odds ratio. The data are as follows:

Diagnosis	Drugs	No Drugs
Schizophrenia	100	12
Affected disorder	12	5
Neurosis	18	24

(b) A study is considered to compare radiation therapy with surgery in treating cancer of the larynx. Use Fisher's exact test to test H_0: θ=1 against H_1: θ≠1.

	Cancer Controlled	Cancer not controlled
Surgery	4	3
Radiation therapy	3	2

(c) Consider a study that examines the relationship between two categorical variables, Gender (Male/Female) and Smoking Status (Smoker/Non-Smoker), in a sample of 500 individuals. The observed counts are as follows:

Gender	Smoker	Non-smoker
Male	60	140
Female	40	260

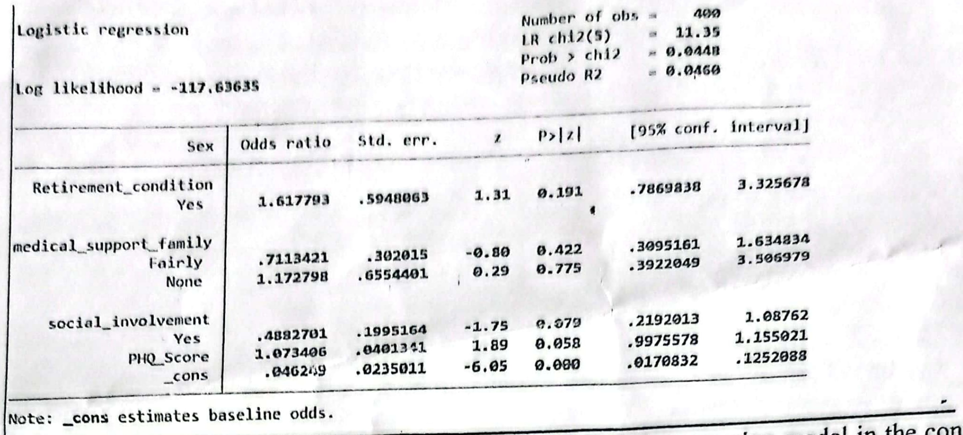
4. (a) What is meant by Generalized Linear Models (GLM)? How do GLMs extend the linear regression framework to accommodate non-continuous response variables? In a GLM, explain why the coefficients are not directly interpretable on the scale of the response variable.

(b) In the context of a binary response variable, compare and contrast the logit, probit, and complementary log-log link functions. Provide insights into when each link function might be preferred.

(c) Explain how you would assess the overall fit of the GLM to the data. Suggest one limitation of using the deviance as a measure of model fit.

5. (a) Provide a detailed explanation of the fundamental concepts of logistic regression. Differentiate it from linear regression and elaborate on how logistic regression models the probability of an event.

(b)



```
Logistic regression                                     Number of obs =        400
                                                        LR chi2(5)    =      11.35
                                                        Prob > chi2   =     0.0448
Log likelihood = -117.63635                             Pseudo R2     =     0.0460

----------------------------------------------------------------------------------------
                     Sex | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------------------+--------------------------------------------------------------
   Retirement_condition  |
                    Yes  |   1.617793   .5948063     1.31   0.191     .7869838    3.325678
                         |
medical_support_family   |
                 Fairly  |   .7113421    .302015    -0.80   0.422     .3095161    1.634834
                   None  |   1.172798   .6554401     0.29   0.775     .3922049    3.506979
                         |
    social_involvement   |
                    Yes  |   .4882701   .1995164    -1.75   0.079     .2192013     1.08762
              PHQ_Score  |   1.073406   .0401341     1.89   0.058     .9975578    1.155021
                  _cons  |    .046249   .0235011    -6.05   0.000     .0170832    .1252088
----------------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Interpret and explain the inference of the results of the logistic regression model in the context of the above output,

6. (a) Define the log-linear model. How to interpret the parameters of the log-linear model and how they can be used to explain joint and conditional associations among variables? In what logistic situation model and log-linear model is appropriate in categorical data analysis?

(b) The following table is taken from a report on the relationship between aspirin use (X) and myocardial infarction (Y) by the Physicians Health Study Research Group at Harvard Medical School. The Physicians Health Study was a five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. Every other day, physicians participating in the study took either one aspirin tablet or a placebo. The study was blind - the physicians in the study did not know which type of pill they were taking.

	Myocardial Infarction	
Group	Yes	No
Placebo	189	10845
Aspirin	104	10933

(i) Fit the independence model and check the goodness of fit. Report $\{\hat{\lambda}_j^Y\}$. Interpret the results of $\{\hat{\lambda}_1^T - \hat{\lambda}_2^T\}$.

(ii) For the saturated model report $\{\hat{\lambda}_{ij}^{XY}\}$. . Show how to interpret these estimates using an odds ratio.

7. (a) Define and explain the concept of a saturated and unsaturated model. Provide an example to illustrate its application in the context of log-linear modeling.

(b) Consider a log-linear model for three-way tables. Calculate the residual degrees of freedom for the below models.

Model	df	G^2
(X, Y, Z)	-	137.93
(XY, Z)	-	131.96
(XY, YZ)	-	7.91
(XZ, Y)	-	120.36
(XYZ)	-	0.00

(c) Test the hypothesis that the expected frequencies satisfy the above-mentioned log-linear model. Select the based model, and explain the reasons behind it.

8. (a) What is matched pair data? Briefly explain different models for matched pair data.

(b) Consider a real-life example of a clinical trial assessing the effectiveness of a new drug treatment for patients with a specific medical condition. Calculate McNemar's statistic.

	Treatment improved	Treatment not improved
Control improved	20	25
Control not improved	30	28

(c) Formulate the logistic regression model for binary matched pair data. Define the key components of the model, including the dependent variable, independent variables, and parameters. Explain how this model can capture the relationship between the treatment and the binary outcome.